Journal of Business Analytics

ISSN: 2573-234X (Print) 2573-2358 (Online) Journal homepage: https://www.tandfonline.com/loi/tjba20

Topic modeling for analyzing open-ended survey responses

Andra-Selina Pietsch & Stefan Lessmann

To cite this article: Andra-Selina Pietsch & Stefan Lessmann (2018) Topic modeling for analyzing open-ended survey responses, Journal of Business Analytics, 1:2, 93-116, DOI: 10.1080/2573234X.2019.1590131

To link to this article: https://doi.org/10.1080/2573234X.2019.1590131

Published online: 16 Apr 2019.

JOURNAL OF BUSINESS ANALYTICS
2018, VOL. 1, NO. 2, 93–116
https://doi.org/10.1080/2573234X.2019.1590131

ORIGINAL ARTICLE

OPEN ACCESS Check for updates

Topic modeling for analyzing open-ended survey responses

Andra-Selina Pietsch[a] and Stefan Lessmann[b]

[a]FactWorks GmbH, Berlin, Germany; [b]School of Business and Economics, Humboldt-University of Berlin, Berlin, Germany

ABSTRACT

Open-ended responses are widely used in market research studies. Processing of such responses requires labour-intensive human coding. This paper focuses on unsupervised topic models and tests their ability to automate the analysis of open-ended responses. Since state-of-the-art topic models struggle with the shortness of open-ended responses, the paper considers three novel short text topic models: Latent Feature Latent Dirichlet Allocation, Biterm Topic Model and Word Network Topic Model. The models are fitted and evaluated on a set of real-world open-ended responses provided by a market research company. Multiple components such as topic coherence and document classification are quantitatively and qualitatively evaluated to appraise whether topic models can replace human coding. The results suggest that topic models are a viable alternative for open-ended response coding. However, their usefulness is limited when a correct one-to-one mapping of responses and topics or the exact topic distribution is needed.

ARTICLE HISTORY
Received 5 November 2018
Revised 14 February 2019
Accepted 27 February 2019

KEYWORDS
Market research; open-ended responses; text analytics; short text topic models

1. Introduction

Surveys are a pivotal research instrument to gain insight into a study subject. In market research, for example, surveys facilitate eliciting the opinions, attitudes, and preferences of consumers and thus provide critical insights for product development and business process management. Open-ended (OE) questions are a crucial component of surveys. They are used to clarify ambiguities and identify opinions that researchers have not thought of before (Lazarsfeld, 1935; Roberts et al., 2014; Schuman, 1966). Likewise, OE questions provide an opportunity to elicit a subject even if a research lacks sufficient knowledge about the topic to define a closed question (Converse, Jean McDonnell, & Presser, 1986). Another advantage of OE questions compared to closed questions is the ability to detect spontaneous thoughts and explore attitudes. Accordingly, common use cases of OE questions in market research include measuring the awareness and recall of brands, attitudes towards a product, or activity as well as likes and dislikes among consumers (Brace, 2018).

However, OE questions also have a major disadvantage: their analysis is associated with high workload. Aiming to identify the topics mentioned in the OE responses and their relative importance, the typical approach requires analysts to read and categorize all or a selection of responses manually (Roberts et al., 2014). Such manual process is time-consuming and prone to errors, especially when multiple researchers analyse the responses separately (between-rater variance) (Tinsley & Weiss, 1975).

The literature suggests several techniques for analysing text data from simple frequency counts (Ten Kleij & Musters, 2003) to advanced machine learning methods (Hong & Davison, 2010; Jin, Liu, Zhao, Yu, & Yang, 2011; Leleu et al., 2011; Mehrotra, Sanner, Buntine, & Xie, 2013; Nguyen, Billingsley, Du, & Johnson, 2015; Phan, Nguyen, & Horiguchi, 2008; Roberts et al., 2014; Weng, Lim, Jiang, & He, 2010; Yan, Guo, Lan, & Cheng, 2013; Zhao et al., 2011; Zuo, Zhao, & Xu, 2016). Text mining OE responses could be a way to circumvent the dilemma between the benefits of having OE questions and the costs associated with their analysis (Roberts et al., 2014). To examine the feasibility of an algorithmic analysis of OE responses, the paper studies unsupervised topic models, which do not require an ex-ante labelling.

Topic models cluster documents based on the assumption that each document is a mixture of latent topics. A quasi-standard in this field is Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). However, LDA is less suitable to process short texts such as OE responses (Sridhar, 2015; Tang, Meng, Nguyen, Mei, & Zhang, 2014). Therefore, the paper consolidates previous work on short text topic modelling and tests the effectiveness of corresponding methods to analyse OE responses in market research.

The short text topic models considered here include Roberts et al. (2014) who implement Structural Topic Models and Leleu et al. (2011) who use Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) to analyse OE responses. Yet, Roberts et al. (2014) have a different focus than the current paper, namely the integration of

CONTACT Stefan Lessmann stefan.lessmann@hu-berlin.de

covariates into topic models, and Leleu et al. (2011) forgo a quantitative and qualitative assessment of the topics although this is essential for the current paper's objectives. Hence, to the best of the authors' knowledge, the literature still lacks a systematic analysis of the potential of topic modelling for OE responses.

Several studies focus on topic extraction from text data that share some characteristics with OE responses, including tweets (Bicalho et al.;2017; Hong & Davison, 2010; Jin et al., 2011; Mehrotra et al., 2013; Nguyen et al., 2015; Weng et al., 2010; Yan et al., 2013; Zhao et al., 2011; Zuo et al., 2016), weblogs (Singh, Waila, Piryani, & Uddin, 2013; Tsai, 2011) and online reviews (Brody & Elhadad, 2010; Titov & McDonald, 2008). Due to the lack of established approaches for OE responses, we examine whether approaches for those three types of corpora can be adapted to OE responses. To shed some light on this matter, Table 1 outlines the most important similarities and differences of OE responses on the one side and tweets, weblogs and online reviews on the other side.

As seen in Table 1, mircoblog entries resemble OE responses in terms of the use of informal language. An important difference concerns the number of covered topics. While tweets usually address a single topic, OE responses often cover multiple ones. The text length is another characteristic where tweets and OE responses display similarities but also differences. Twitter enforces a maximum length of 140 characters per tweet. Market research surveys do not enforce a maximum length for OE responses so that these can be substantially longer. In practice, however, survey respondents often provide only short answers to OE responses. For example, Gendall, Menelaou, and Brennan (1996) report an average response length between 4.5 and 7 words per response. These figures are consistent with the experience of the market research agency that supports the focal study through providing real-world data. As detailed in Section 3.1, the data we employ exhibits an average length of 5.5 words per OE response. In this regard, we suggest that the length of tweets and OE responses is, in practice, often similar on average whereby the length of OE responses exhibits much larger variance than that of tweets. This also suggests that microblog entries are more similar to OE responses than weblog entries and online reviews, which share the language style but differ in document length.

The shortness of OE responses, which is often observed in practice, represents the main challenge for topic modelling in the market research context considered in this study. As microblog entries and OE responses resemble each other in terms of length (Naveed, Gottron, Kunegis, & Alhadi, 2011), a brief overview of related work with a focus on topic modelling for short text, mostly applied to tweets, is provided in the following.

Several techniques for extracting topics from short texts have been proposed in the literature. A recent study of Bicalho et al. (2017) systematizes the field and introduces a general framework for overcoming the specific challenges of short text topic modelling. In general, short text topic models split into two categories: The first one uses auxiliary information to enrich the input (knowledge-based approaches). Examples include corpus-related metadata (Hong & Davison, 2010; Mehrotra et al., 2013; Weng et al., 2010), external knowledge sources like auxiliary long text (Jin et al., 2011; Phan et al., 2008) or word embeddings (Bicalho et al., 2017; Nguyen et al., 2015). The second category includes corpus-based approaches that rely exclusively on the target corpus, meaning the text corpus from which topics shall be extracted; such as the collection of OE responses in this paper. Corpus-based approaches modify the topic modelling process itself (Mihalcea, Courtney, & Strapparava, 2006). Examples include the introduction of stronger assumptions about the data (Bicalho et al., 2017; Nguyen et al., 2015; Zhao et al., 2011) or the manipulation of the document generation process (Yan et al., 2013; Zuo et al., 2016). Table 2 outlines relevant prior studies, divided into knowledge-based and corpus-based approaches, including the respective target corpora and methodology. It further shows where to localize the current study, which fills the gap of short text topic models applied to OE responses in both categories.

Using a set of real-world OE responses from a market research company, this study explores the potential of three short text topic models for OE responses and compares them to LDA as a benchmark: Latent Feature LDA (LFLDA) (Nguyen et al., 2015), Biterm Topic Model (BTM) (Yan et al., 2013) and Word Network Topic Model (WNTM) (Zuo et al., 2016). In each of the three

Table 1. Comparison of different types of data with OE responses.

Data	Similarities with OE responses	Differences from OE responses
Microblog entries (e.g., tweets)	• Document shortness, informal language (Naveed et al., 2011) • While OE Responses can be much longer than tweets, survey respondents often provide only relatively short answers of 4.5 to 7 words on average (Gendall et al., 1996)	• Coverage of a single topic (Zhao et al., 2011) • Coverage of broad topics like politics or sports (Hong & Davison, 2010; G. Lockot, personal communication, September, 2017)
Weblog entries	• Informal language	• Document length (Singh et al., 2013)
Online reviews	• Informal language • Topic granularity (focus on specific details) (Liu, 2012)	• Document length

Table 2. Exemplary research on topic modelling for short text.

Approach	Authors	Target corpus	Methodology
Knowledge-based	Hong and Davison (2010), Mehrotra et al. (2013), Weng et al. (2010)	Tweets	Aggregation of short documents to longer pseudo documents based on metadata
	Jin et al. (2011), Phan et al. (2008)	Web search snippets, advertisement, tweets	Topic modelling on external long text (e.g., Wikipedia)
	Nguyen et al. (2015)	News titles, tweets	Incorporation of word vectors trained on large corpora (e.g., Google news) (LFLDA)
	Bicalho et al. (2017)	Tweets, news articles, news titles, web search snippets	Distributed Representation-based Expansion (DREx): Generate longer pseudo-documents based on word vectors
	This study	*OE responses*	*Incorporation of word vectors trained on large corpora (LFLDA)*
Corpus-based	Nguyen et al. (2015), Zhao et al. (2011)	News titles, tweets	Restriction of one topic per document
	Yan et al. (2013)	Tweets	Modelling topic distributions for biterms (BTM)
	Zuo et al. (2016)	Weibo entries	Modelling topic distributions for words (WNTM)
	Bicalho et al. (2017)	Tweets, news articles, news titles, web search snippets	Co-Frequency Expansion (CoFE): Generate longer pseudo-documents based on word co-occurrence
	This study	*OE responses*	*Modelling topic distributions for biterms (BTM) and words (WNTM)*

studies, the proposed short text topic modelling approach has been compared to LDA as baseline using data related to microblog entries. The studies consistently observe an improvement over this baseline suggesting that all three methods outperform LDA on microblog entries. WNTM additionally shows good performance when dealing with topic imbalance (Zuo et al., 2016). This is relevant for OE responses as usually some topics are mentioned much more frequently than others. Further, the methods are not associated with any assumptions or requirements that are not transferable to OE responses, like the restriction of having only one topic per document or the need for metadata. Hence, we consider their potential for analysing OE responses as high.

Table 2 suggests that the extraction of topics from short texts has received considerable attention in previous work. However, we also observe from Table 2, that corresponding studies have not looked into the specific application context of OE responses, which is the goal of this paper. Using real-world data from user surveys, we add to the literature by providing original empirical evidence concerning the potential of selected short text topic models in OE response processing. More specifically, the paper makes two contributions: First, it investigates the extent to which topic modelling can replace manual analysis of OE responses. To that end, we evaluate topic model results along two dimensions: the comprehensibility of extracted topics (topic quality), and the amount of information to represent OE responses and derive the topic distribution (topical document representation). Both dimensions are relevant for the suitability of topic modelling in market research. Second, the paper elaborates on the relative merits and demerits of alternative short text topic models to provide guidance for researchers and practitioners how to choose the right method for a given market research task.

2. Methodology

2.1. Latent Dirichlet allocation

Topic modelling is an approach to cluster text documents, assuming that each document is a function of latent variables called topics (Aggarwal & Zhai, 2012). LDA, introduced by Blei et al. (2003), represents a state-of-the-art method in this field (Hong & Davison, 2010). Yet, despite its wide popularity, LDA does not work well for every kind of text data. While it successfully models topics for corpora like news articles (Blei et al., 2003) and scientific papers (Griffiths & Steyvers, 2001), it shows disappointing results for short documents and small corpora[1] (Sridhar, 2015; Tang et al., 2014). In the latter cases, data sparsity and limited context prevent a reliable extraction of document-based word co-occurrences, which is the basis for LDA (Sridhar, 2015). Also, LDA tends to detect frequent topics better than rare ones (Zuo et al., 2016) and broad topics better than specific ones (Titov & McDonald, 2008). Thus, corpora with imbalanced topic distributions and those that require a detailed analysis are also challenging. These critical characteristics apply to OE, which leads to the assumption that LDA is not ideal for this kind of data. LDA serves as benchmark in the empirical part of the paper and foundation to introduce short text topic models.

LDA is a three-level hierarchical Bayesian model where each document d_m is modelled as a finite mixture over a set of K corpus-wide topics z_k (Blei et al., 2003). Each topic, in turn, is a distribution over a fixed set of V words w_v. As a generative model, LDA assumes that the words that a document contains are generated by the latent topics. Therefore, LDA tries to infer the latent topics that could have generated the documents. For finding these topics, LDA uses the word co-occurrence pattern in the corpus, which is withdrawn from the document-

term matrix (DTM). In doing so, a key component of LDA is the "bag-of-words" assumption, meaning that the order of words is ignored (Blei et al., 2003). The more often two words co-occur in a document, the more likely they belong to the same topic (Aggarwal & Zhai, 2012).

The generation process can be formally described as follows (Blei et al., 2003):

(1) For each topic z, choose the probabilities over words $\phi_z \sim Dir(\beta)$, where ϕ_z is drawn from a symmetric Dirichlet prior distribution with parameter β.
(2) For each document d, choose the probabilities over topics $\theta_d \sim Dir(\alpha)$, where θ_d is drawn from a symmetric Dirichlet prior distribution with parameter α.
(3) For each word w_{dn} in document d, choose a topic $z_{dn} \sim Multinomial(\theta_d)$ and then choose a word w_{dn} from the multinomial distribution $w_{dn} \sim Multinomial\left(\phi_{z_{dn}}\right)$.

The functioning of LDA is often illustrated using the plate notation of Figure 1 where a circle represents a random variable and an arrow a unilateral dependency between variables. The processes within a box are repeated multiple times with capital letters giving the number of repetitions.

The number of topics K as well as the Dirichlet hyperparameters α and β are determined prior to modelling. The parameter α denotes the prior document-topic distribution and the parameter β the prior topic-word distribution (Griffiths & Steyvers, 2001). The posterior distributions of θ_d, ϕ_z and z are inferred by using collapsed Gibbs sampling (Griffiths & Steyvers, 2002), following previous works (Griffiths & Steyvers, 2001; Nguyen et al., 2015; Yan et al., 2013; Zuo et al., 2016).

2.2. Application of topic models to open-ended responses

Market researchers are mainly interested in two things: Identifying the topics that are mentioned in OE responses and the topics' relative distribution. The former is provided by the posterior topic-word distribution ϕ, which is one output of a topic model. ϕ provides the likelihood for each word belonging to each topic. By considering only the top words, i.e. those that are most likely to appear in a topic, one can derive the content of the topics (Blei et al., 2003). The top words are most interesting because the lower the topic-word probability, the weaker the topic-word relation. Topic models do not provide labels for the topics so that the interpretation and labelling of extracted topics is left to the researcher (Schouten & Frasincar, 2016).

The posterior document-topic distribution θ_d can provide insights into the topics in addition to the top words. θ_d is represented as a $M \times K$ matrix where for each document d and each topic z, the probability $P(z|d)$ shows how likely it is that z is present in d. θ_d can be used to find the most representative documents (top documents) for z, i.e. the documents with the highest document-topic probability for z. The top documents can help to further describe a topic (Aggarwal & Zhai, 2012).

The share of documents that contain a topic compared to the corpus size can also be derived from θ_d. By choosing a threshold t, one can assign only those topics to each document for which $P(z|d) > t$. This can be used to compute the share of the topics over the whole corpus. In market research, the share of documents corresponds to the share of respondents mentioning a certain topic.

2.3. Short text topic models

This section introduces the three short text topic models LFLDA, BTM and WNTM. It briefly presents the differences to LDA and explains why they are more suitable for OE responses.

2.4. LFLDA

Nguyen et al. (2015) complement the sparse co-occurrence pattern in short documents through integrating vector representations of words (hereinafter: word vectors). They use two sets of pre-trained word

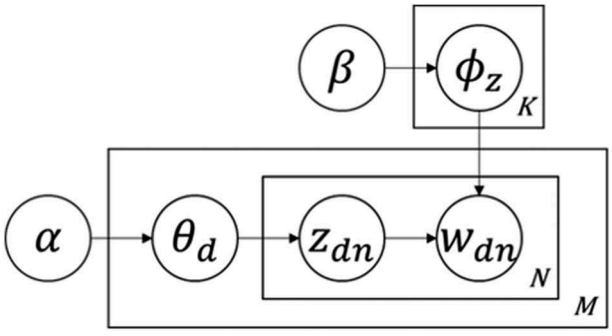

Figure 1. Graphical representation of the generative process of LDA. Adapted from (Aggarwal & Zhai, 2012; Blei et al., 2003).

vectors: The first one is trained on a subset of the Google News corpus via Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013) and the second one on Common Crawl web data via Global Vectors for Word Representation (GloVe) (Pennington, Socher, & Manning, 2014).

As for LDA, LFLDA tries to find the latent topic structure that could have generated the observed documents. The generative process is similar to LDA but differs in the way how words are generated from topics. In LDA, a word can only be drawn from the Dirichlet multinomial distribution ϕ that is trained on the target corpus. In contrast, LFLDA allows each word to be drawn from either that distribution or from a multinomial distribution based on the vector representation of every word and topic in the corpus. By incorporating the vector representations, LFLDA uses information about word-topic relations from larger external corpora. Hence, LFLDA circumvents the issue of LDA with the sparse information in short text about the word co-occurrence structure.

To determine from which of the two distributions a word w_{dn} is drawn, a binary indicator variable s_{dn} is sampled from a Bernoulli distribution $Ber(\lambda)$. The hyperparameter λ determines the probability with which a word is sampled from the latent feature component.

2.5. BTM

In contrast to LFLDA, BTM (Yan et al., 2013) does not use an external knowledge source to deal with the short documents' missing context. However, it differs from LDA in two other regards that concern the topic modelling input and the generative process.

First, the input to topic modelling is not the set of documents D as in LDA but the corpus-wide set of biterms B. A biterm b is defined as "an unordered word-pair co-occurred in a short context" (Yan et al., 2013, p. 1446) where a short context denotes a document. For example, the document "great customer service" consists of three biterms: "great customer", "customer service" and "great service". The biterm approach of LFLDA bases on the assumption that there is a topic distribution θ for the entire corpus instead of a topic distribution θ_d for each document. Consequently, the hyperparameter α denotes the prior corpus-topic distribution and not the document-topic distribution.[2]

Second, LDA uses the word co-occurrence pattern per document to generate words. In contrast, BTM generates biterms instead of single words. The aim of the generative process in BTM is finding the latent topics that could have generated the biterms, which make up the corpus.

As the topic inference in LDA is based on the word co-occurrences per document, the issue with short text like OE responses is that their shortness leads to a relatively sparse word co-occurrence structure per document. The major advantage of BTM is that it uses the entire corpus as input, which makes the topic model insensitive to document shortness and hence improves the detection of topic-word relations.

2.6. WNTM

WNTM (Zuo et al., 2016) infers topic distributions for words instead of documents to circumvent the sensitivity of LDA towards document length. This requires a transformation of the input documents. By moving a sliding window of length S through each document, a word co-occurrence network is created where the network nodes represent the vocabulary of the corpus and the edges the co-occurrences of each word pair weighted by the number of co-occurrences in the corpus. Subsequently, for each word w_v a pseudo-document d^p is created that consists of all words that co-occur with w_v, i.e. all words that are connected to w_v in the word network. Instead of using the original text documents as input to topic modelling, as done in LDA and LFLDA, the newly generated pseudo-documents are used as input in WNTM. Hence, the key difference between the generative processes of LDA and WNTM is that WNTM does not generate the original but the pseudo-documents.

The key difference between the output of LDA and WNTM is the interpretation of θ_{d^p} which denotes the probability of each topic being present in a pseudo-document d^p. A pseudo-document entails a word's context information across the entire corpus. Hence, θ_{d^p} is regarded as the distribution over topics for each word, where each word in turn is represented by its pseudo-document.

The advantage of using WNTM for short text like OE responses is twofold. First, modelling topics for words by considering a word's co-occurrences across the entire corpus decreases the model's problem with document shortness. Similar to BTM, this improves topic extraction as the words' contextual information are not limited to the co-occurrences within a -document. Second, there are more words than documents that are related to rare topics. Thus, the authors claim that WNTM is better capable of detecting rare topics than other topic modelling approaches (Zuo et al., 2016). This is relevant for OE responses as usually some topics are mentioned by much more respondents than others.

3. Experimental design

3.1. Data

To examine whether topic modelling can serve as an alternative for analysing OE responses and which of

the selected topic models works best for this kind of data, several experiments are conducted on real-world OE responses. The data source and pre-processing tasks as well as a summarization of the corpus' main characteristics are presented in the following.

3.1.1. Data source

The dataset is provided by a Berlin-based market research company (hereinafter: partner company). The data belongs to an online survey of software developers, which is repeated quarterly. The current paper focuses on an OE question of this survey where developers are asked why they recommend developing on a certain platform. The data was gathered between December 2014 and July 2017 and 7,743 responses are available for this question. This set of responses makes up the target corpus for this paper.

Each quarterly repetition of the study is analysed separately by the partner company. Because of the high workload that is associated with the evaluation of OE responses, only a random sample of approximately 450 responses per wave is manually coded. This leads to 5,001 labelled responses in total. There are nine different labels that can be assigned to the responses. Responses that cannot be assigned to any of those labels are classified as "other". This "other" category is a collection of side issues deemed too small to get an own label. A team of researchers is responsible for coding, some of whom have been involved in the project from the start while others were only involved in some waves. In total, seven researchers have been involved in the coding (G. Lockot, personal communication, September, 2017).

3.1.2. Pre-processing

Several pre-processing steps are conducted to increase the quality of the dataset and to transform data in such a way that it complies with the requirement of (short text) topic models. First, standard pre-processing tasks are performed, including the translation of non-English responses, lemmatization, conversion to lowercase and the removal of numbers, punctuation, stop words and infrequent words (Manning, Raghavan, & Schütze, 2009). This leads to a vocabulary of $V = 766$ unique words and a corpus of $M = 7,622$ documents.

For LFLDA, BTM and WNTM some method[3] - specific data preparation is performed. For LFLDA, a set of pre-trained GloVe word vectors (Pennington et al., 2014) is chosen following (Nguyen et al., 2015). The set is trained on 42 billion tokens of Common Crawl web data and contains 300-dimensional vectors for 1.9 million words.[4] For BTM, all documents shorter than two words are excluded from model training, which leaves 6,993 documents. Similarly for WNTM, all documents shorter than the window size S are excluded from topic modelling. By setting S = 3 in this work, the ratio between average document length and window size is similar to the one used in the original work by Zuo et al. (2016). This leads to 5,776 documents for model training. Later, topics can also be inferred for the documents that are excluded from model training in BTM and WNTM.

3.1.3. Descriptive analysis

To get a clearer picture of the data, several descriptive analyses are conducted. After pre-processing the documents contain between one and 160 words with an average length of 5.5 words while 75% of the responses contain seven words or less. Recall that these values can differ for certain short text topic models due to model-specific preprocessing. For example, the minimum word length per response will be two and three for BTM and WNTM, respectively. In general, one may question the minimum and maximum number of words per response. For example, a text of 160 words may not be regarded as short anymore; after all it is much longer than a tweet. In this study, we do not enforce pre-defined thresholds, unless required by a specific topic modelling method. Rather, we employ common text pre-processing techniques and proceed with the resulting document lengths. Given the scarcity of prior work dedicated to topic modelling from OE responses, we suggest that the application of a text standard pre-processing pipeline is suitable for this paper. Enforcing overall limits of the minimum and maximum number of words per response would require a systematic approach to set these limits. Developing corresponding methodology is a valuable goal for future research but beyond the scope of this paper, which seeks insight into the relative suitability of available short text topic models for OE response processing.

Aside from the document length, the distribution of the manual labels is of interest as they serve as a gold standard for the evaluation in this study. The pre-processed corpus includes 4,958 labelled documents for all methods. Most documents are assigned to only one label but there is also a significant share of documents with multiple labels (Figure 2). This supports the assumption that topic models that allow only one topic per document – as for instance used in Zhao et al. (2011) for tweets – are not suitable for OE responses. Aside from the number of labels per response, the overall importance of each label is relevant. The set of labelled responses shows an imbalanced label distribution, i.e. the share of responses assigned to each label differs significantly as depicted in Figure 3.[5] It means that there are substantially more documents that provide information about some labels than others. Appendix A provides short descriptions of the labels.

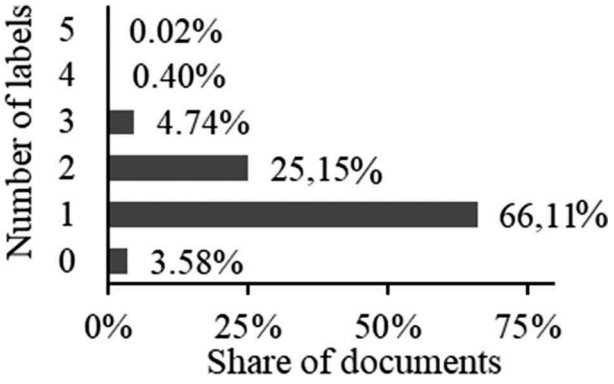

Figure 2. Number of labels assigned to each document.

The responses with zero labels are not unlabeled responses. Here, the researchers decided that they could not assign the responses to any of the nine labels. So, they assigned them to the previously mentioned "other" category.

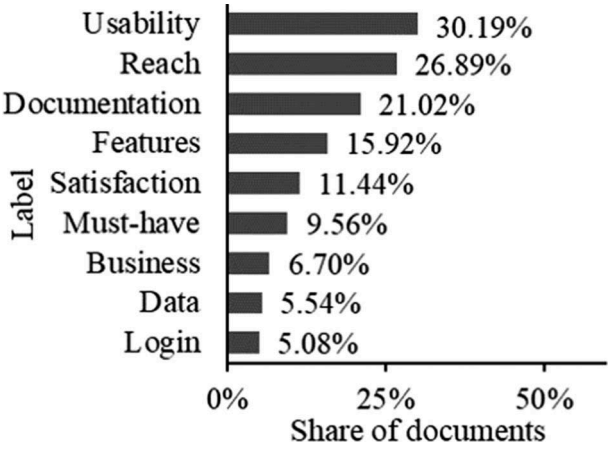

Figure 3. Share of documents assigned to each label.

3.2. Model implementation

The three short text topic models and LDA as benchmark are implemented using R, Python, Java, C++ and Bash. The detailed technical specification of the infrastructure employed for data pre-processing, model fitting and evaluation is as follows: of a personal computer with Intel i7-6500U CPU, running on Windows 10 with R version 3.4.2, Java Development Kit version 1.7 and Python version 3.5. LDA is trained using the R package *topicmodels* (Hornik & Grün, 2011). For the other three methods, published source code13F[6] is used and adapted to the present application (e.g., hyperparameter settings and evaluation).

For each method, different hyperparameter settings are evaluated. Some authors (Lu, Mei, & Zhai, 2011; Yin & Wang, 2014) suggest smaller values for α within conventional LDA when applied to short text to improve performance compared to the common setting of $\alpha = 50/K$. For instance, Yan et al. (2013) use $\alpha = 0.05$ and use Nguyen et al. (2015) $\alpha = 0.1$ when using LDA for short text. Moreover, Tang et al. (2014) propose smaller values for β when dealing with short text, for example $\beta = 0.01$ as set in Nguyen et al. (2015) and Yan et al. (2013). Therefore, it is assumed that rather small values for α and β are appropriate in this work. This implies that documents are associated with rather few topics (small α) and that topics are rather word-sparse and thus better to distinguish from each other (small β). Guided by the parameter settings with the best performance in the original papers (Nguyen et al., 2015; Yan et al., 2013; Zuo et al., 2016), two values for each of the hyperparameters are implemented. For reason of comparability, the values for α (for BTM: α^B) and β are identical for all methods. Moreover, for each method, the number of topics K is varied from five to 50 with a step size of five. As the number of topics mentioned by respondents can change for different studies, this variation is important to understand how the models behave when K is small or large. The range for K is chosen based on the manual labels given. The lower boundary is very close to the original number of labels. Meanwhile, the upper boundary is a trade-off between a sufficiently large value to observe a trend based on varying K while sustaining the feasibility of a manual inspection of topics. Table 3 summarizes the hyperparameter settings and the resulting number of models trained per method. This amounts to 200 models in total. Parameter inference is done via Gibbs sampling with 1,000 iterations for all models. Finally, Figure 4

Table 3. Model parameters and number of models trained.

Method	Hyperparameters	Number of topics	Number of models trained
LDA	$\alpha = \{0.05,\ 0.1\} \beta = \{0.01,\ 0.1\}$	$K = \{5, 10, \ldots, 45, 50\}$	40
LFLDA	$\alpha,\ \beta$ (as LDA) $\lambda = \{0.6,\ 1\}$	K (as LDA)	80
BTM	$\alpha^B,\ \beta$ (as LDA)	K (as LDA)	40
WNTM	$\alpha,\ \beta$ (as LDA)	K (as LDA)	40

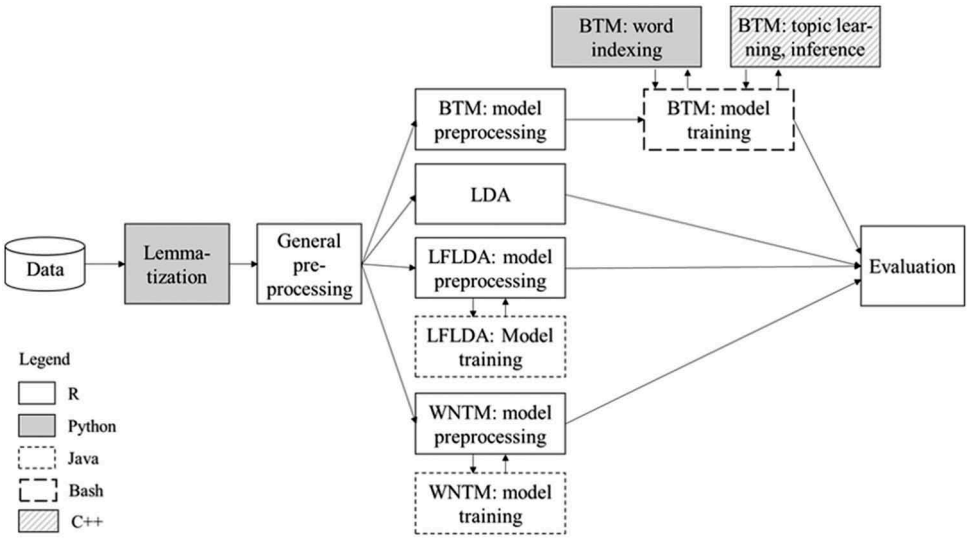

Figure 4. Architecture of the empirical analysis.

summarizes the overall architecture of the experiments.

3.3. Performance measurement

Lau, Newman, and Baldwin (2014) and Chang, Boyd-Graber, Gerrish, Wang, and Blei (2009) suggest that topic models have two main use cases, direct human consumption and text preparation. The former case entails a manual analysis of extracted topics to interpret their meaning while in the latter case another text processing algorithm, for example a text classifier, operates on the basis of the extracted topics. In this paper, both perspectives are relevant.

First, the topics must be sufficiently clear for exploratory purposes (in the following referred to as quality of topics). A statistically reasonable topic is not necessarily regarded as meaningful by a human (Newman, Karimi, & Cavedon, 2009). Some topics (e.g., "advertisement, targeting, audience, viral, brand") may be perceived as more interpretable than others (e.g., "company, time, easy, app, tools"). A common approach is to evaluate the quality of topics by considering its top ten words, i.e., the ten words that are most likely to be drawn from that topics (Newman, Lau, Grieser, & Baldwin, 2010). This procedure is also used here.

Second, the topics need to contain enough information to represent the documents appropriately (in the following referred to as topical document representation). This is required to deduce the topic distribution, i.e., the share of responses mentioning each topic. It is common practice to evaluate the topical document representation based on the performance of topic models on extrinsic tasks like document clustering or classification (Blei et al., 2003; Nguyen et al., 2015; Yan et al., 2013; Zuo et al., 2016).

Both dimensions – quality of topics and topical document representation – are evaluated in this paper using a quantitative as well as a qualitative approach for each. The quantitative approaches make it possible to objectively compare the topic modelling methods. Meanwhile, the qualitative approaches complement the quantitative evaluation by gaining a deeper insight into some selected examples of topics or topic models. The latter also allows to integrate expert knowledge. Table 4 summarizes how the model evaluation will be conducted on the four dimensions.

The dual evaluation approach of assessing extracted topics from a quantitative and qualitative angle is beneficial to obtain a comprehensive picture of the potential of short text topic models. However, the evaluation approach also has implications that need to be acknowledged. On the one hand, the quantitative assessment requires OE responses to have undergone manual labelling. The assessment then translates into comparing manual to algorithmically generated labels. The qualitative evaluation, on the other hand, requires the involvement of market research experts to judge extracted topics and compare the outputs of different short text topic models

Table 4. Performance measurement on four dimensions.

		FOCUS	
		Quality of quality	Topical document representation
APPROACH	**Quantitative**	Goal: Compare all topic models with regards to topic quality Metric: Coherence score by Mimno et al. (Mimno, Wallach, Talley, Leenders, & McCallum, 2011) Calculation: Compute a coherence score per topic by using its top word list and average over all topics to get a single coherence score per topic model (Lau et al., 2014) (Implementation with R package SpeedReader (Denny, 2017)); the closer the score to zero, the higher the indicated coherence Benefits: No external information needed, high correlation with human judgement (Lau et al., 2014; Mimno et al., 2011) References: (Yan et al., 2013; Zuo et al., 2016)	Goal: Compare all topic models with regards to topical document representation Metric: F1 score for document classification with Support Vector Machines (SVM) (Manning et al., 2009; Van Rijsbergen, 1979) Calculation: Fit a binary classification task for each of the nine labels (dependent variable) where the document-topic probabilities θ_d are the independent variables (Manevitz & Yousef, 2001), using SVM as a classifier (Implementation with the R package caret (Kuhn, 2008); calculate performance metric F1 score per classification task and average over all tasks to get a single metric per topic model (Manning et al., 2009) Benefits: Metric is common in information retrieval (Van Rijsbergen, 1979), SVM have shown to be effective in text classification (Manning et al., 2009) References: (Blei et al., 2003; Nguyen et al., 2015; Yan et al., 2013; Zuo et al., 2016)
	Qualitative	Goal: Understand the usefulness of exemplary topics by leveraging expert knowledge Procedure: Two experts from the partner company independently interpret eight topics (two topics per method), label them and compare them to each other without knowing which topic is produced by which method	Goal: Investigate if the topic distribution of exemplary topic models on a corpus-level is a good approximation to the distribution of the manual labels (Figure 3) Procedure: First, for $K = 10$ and $K = 20$, the topic models with the best quantitative performance are chosen for further investigation (these values of K are chosen together with the experts as $K = 10$ is close to the original number of labels and $K = 20$ approximately represents the number of sub labels the experts see in the data; this is to see how K affects the performance on topical document representation). Then, for both topic models, the topics are matched with the manual labels and a topic z is assigned to a document d if the document-topic probability is larger than a threshold t (using different values for t); based on this allocation, the topic distribution is calculated and compared to the label distribution

to one another. Therefore, the quantitative and qualitative evaluation both enforce sharp constraints on the type and amount of data that can possibly be considered in the study. As explained above, we have access to roughly 5,000 OE responses gathered from a recurring survey between December 2014 and July 2017. Expanding the amount of data were desirable but is prohibited by the strict requirements of the evaluation approach. This also implies that research findings and conclusions are limited to the specific type of OE responses employed in the study while a replication of the empirical analysis to test external validity is left to future research.

4. Results

4.1. Quality of topics – quantitative evaluation

For each value of K, four models are trained for LDA, BTM and WNTM each using different hyperparameter combinations of α and β. For LFLDA, eight models are trained, as this method additionally includes the hyperparameter λ, for which also two values are used.

Figure 5 gives an overview of the coherence scores produced for the different methods. The closer the coherence score to zero, the higher the topic coherence averaged over all topics produced by a topic model. The scores for all trained models are reported in Appendix B. The lines in Figure 5 show the best scores reached by each method across all hyperparameter settings. These show that no method significantly outperforms the others for $K \leq 10$. In contrast, for $K \geq 15$, BTM achieves the highest scores and its advantage increases with K.

Yet, the lines only show the *best* coherence scores produced by each method. To examine if the superiority of BTM depends on a certain hyperparameter setting, the shaded areas in Figure 5 depict the ranges of scores per method that are produced by the different parameter settings. The boundaries of the shaded areas equal the scores for the best (upper boundary) and the worst parameter combination (lower boundary) for each K. The figure shows that the performance of BTM is less sensitive to different parameter settings compared to the other methods, meaning that the coherence scores achieved by the best and the worst models differ less. Yet, it must be noted that twice as many hyperparameter settings are implemented for LFLDA, which limits the comparability to the other methods' ranges. However, there is no hyperparameter combination that consistently produces the best results for any method (Appendix B).

Another interesting observation is the downward trend of all methods' scores with an increasing number of topics. One possible reason is that all topics are generally worse when K is high. Another explanation could be that there are still good topics but as there is only a limited number of topics in the corpus, increasing the value of K leads to more nonsense topics with very low coherence scores. Eventually, this decreases average coherence scores. To investigate this, Figure 6 depicts for every method and every K the scores of the most and least coherent topics over all models. Notably, the best scores produced by all methods show no dependence on the number of topics. This means that regardless of the value of K, there is still at least one relatively good topic. In contrast, the scores of the least coherent topics decrease remarkably with K. Both observations indicate that topic models with a high number of topics still produce good topics but the larger K, the more incoherent topics are produced which decreases the average scores.

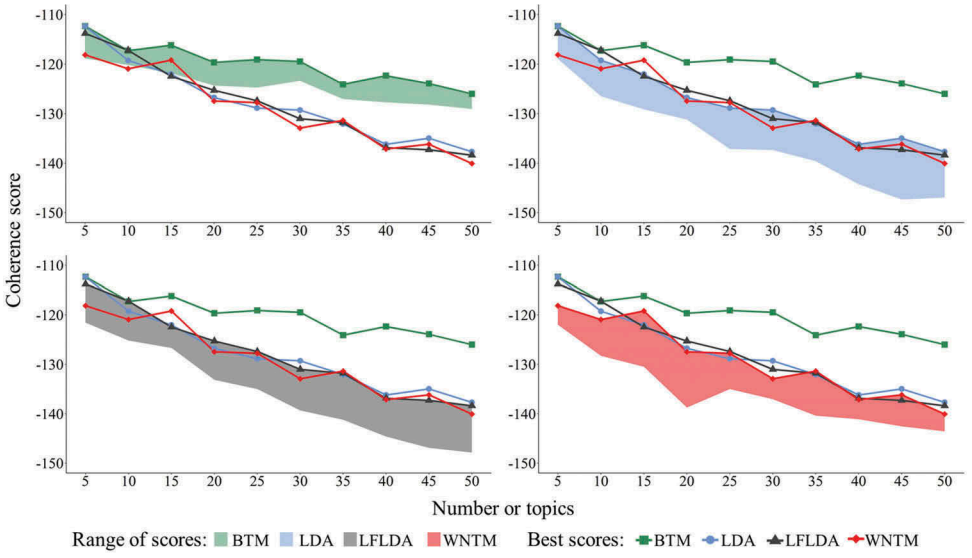

Figure 5. Best average coherence scores per method (lines) and range of average coherence scores per method produced by different hyperparameter combinations (shaded area).

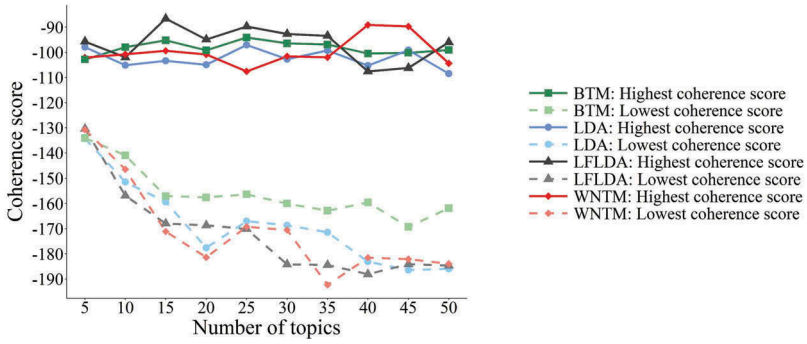

Figure 6. Highest and lowest coherence scores per method on the topic-level.

In summary, the quantitative evaluation of topic coherence indicates that BTM produces on average more coherent topics regardless of the hyperparameter setting. Apart from that, it is hard to recognize a difference between LDA, LFLDA and WNTM. For some values of K, LDA even outperforms LFLDA and WNTM although the differences are comparatively small. Moreover, the results show that the different numbers of topics reveal valuable insights since K influences the model ranking as well as the absolute coherence scores.

4.2. Quality of topics – qualitative evaluation

This section explores the topic interpretability from a qualitative perspective. To achieve this, the opinions of two domain experts are used and compared to the quantitative coherence scores. Only the models for $K = 20$ are considered for the qualitative evaluation. This value is chosen based on two criteria: First, it is relatively close to the number of original labels, which is nine.

This increases the likelihood that the topic granularity is similar to the one the experts are used to. Second, as seen in Figure 5, BTM increasingly deviates from the other methods when K increases. For $K = 20$, there is already a notable distance between the score of BTM and the remaining methods. This helps to examine whether the experts' perception of differences in topic coherence is consistent with the quantitative scores. For each method and $K = 20$, the model with the highest average coherence score is considered. These are also the ones depicted by the lines plotted in Figure 5.

Table 5 shows the eight topics and their coherence scores, which are evaluated by the two experts. The word lists are ordered by topic-word probability, i.e., the first word in each list is most likely and the last word least likely to be generated by the respective topic. Many words appear in every method (e.g., "easy" for topic A) but only few words are unique to one method. Further, the unique words are rather positioned at the end of the lists, meaning that the topics are even more similar when focusing only on the top words. Regarding the coherence scores, there is another interesting finding: The least coherent topic in the table is topic B of LFLDA and the most coherent one is topic B of BTM. However, both topics contain seven identical words in the beginning and only differ in the ordering and the last three words.

The evaluation through the experts happens separately but their opinions hardly differ. First, both state that all topics are generally understandable. Regardless of the methods, they interpret the topics as follow: Topic A is about good documentation and user-friendliness and topic B about the large user base of the platform. Both regard topic B as more coherent and useful than topic A because they see two separate themes – documentation and user-friendliness – in topic A, which from their perspectives should belong to two separate topics. Meanwhile, topic B covers only a single topic and is therefore regarded as more coherent. This is not in line with the coherence scores, which indicate a higher coherence for topic A for LDA, LFLDA and WNTM and very similar scores for BTM. Moreover, one expert highlights the last two words

Table 5. Top words and coherence scores for two exemplary topics per method.

Method (Topic)	Score	Top words (underlined words appear in all methods and italic words are unique to one method)
BTM (A)	−102.24	good, documentation, easy, api, sdk, pretty, use, work, platform, *user*
LDA (A)	−124.76	well, document, easy, api, use, work, sdk, simple, pretty, *integrate*
LFLDA (A)	−124.55	well, document, easy, api, use, work, simple, *quite, clear*, sdk
WNTM (A)	−107.33	easy, api, use, well, document, documentation, simple, good, work, platform
BTM (B)	−101.02	user, reach, audience, large, platform, base, huge, use, *good, easy*
LDA (B)	−128.66	reach, people, use, *lot*, platform, audience, *many*, user, *can, way*
LFLDA (B)	−129.06	user, base, large, audience, huge, reach, platform, potential, *big*, wide
WNTM (B)	−128.45	user, reach, audience, base, large, huge, people, platform, potential, wide

of topic B of LDA ("can", "way") which he regards as confusing in this context. In contrast, he likes the words "potential" and "wide" within LFLDA and WNTM and thinks they make the topic even clearer. This is again inconsistent with the coherence scores that indicate a higher coherence for LDA than for LFLDA. For topic A, one expert expresses a slight preference for LDA and the other one for LDA and LFLDA. However, they call it rather a gut feeling than a reasoned decision. For topic B, they state that the topics except for LDA are so similar that they cannot name a preference between BTM, LFLDA and WNTM.[7]

To compare the topics, the experts mainly focus on the last words in the lists although these are less representative for the topics than the first words. However, the experts' approach is understandable because the last words are those that differentiate the methods from each other. It can be questioned whether the order in which the words appear in the topics really matters or if the words are more or less equally likely to be drawn from the topics. To investigate this based on an example, the topic-word distributions for topic B for BTM and LDA are explored. These two topics are of special interest regarding their last words as mentioned above: First, topic B of BTM achieves a notably higher coherence score than LFLDA although it differs only in the last three words. Second, one expert highlights the inappropriateness of the last two words of topic B of LDA "can" and "way". Figures 7 and 8 show for both topics that the words at the beginning of the lists are significantly more likely to be drawn from a topic than those at the end of the lists. Surely, a comparison of the topic-word distributions for all topics would allow a more complete and generalizable interpretation. But the two examples already show that one should be careful when putting too much weight on the last terms in the top word lists.

In summary, the qualitative evaluation shows that experts who are familiar with OE response coding regard the exemplary topics as interpretable. Further, the results imply that the qualitative evaluation is not always in line with the quantitative

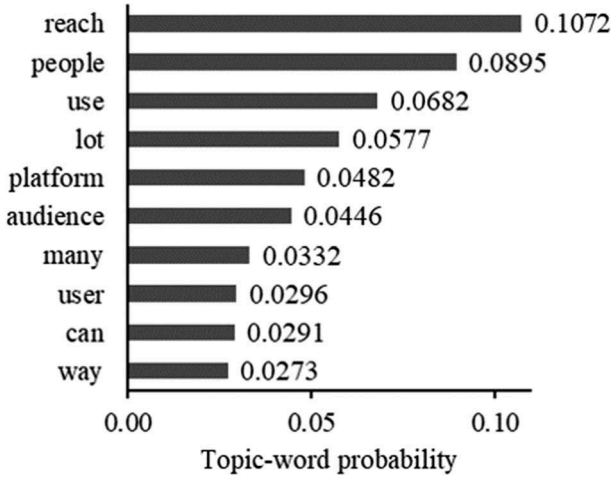

Figure 7. Topic-word probabilities for the exemplary topic B (LDA).

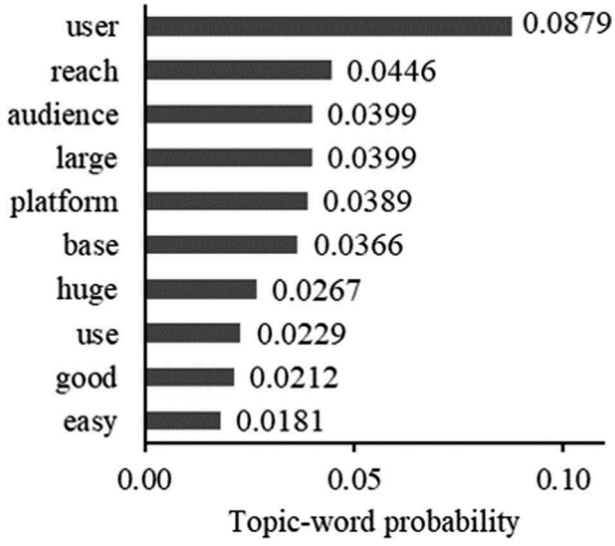

Figure 8. Topic-word probabilities for the exemplary topic B (BTM).

coherence score. For instance, the clear superiority of BTM reflected in the quantitative scores is not reproduced by the expert judgements. Although it is not the purpose of this section to prove or disprove the reliability of the coherence score, previous results suggest that one should not have blind faith in it. Moreover, the investigation of the topic-word probabilities implies another interesting finding. Although it is common practice to look at top ten words lists when interpreting topics (Newman et al., 2010), one should maybe rethink approaches for topic visualization. As seen in Figures 7 and 8, the first terms in the top word lists should be weighted stronger than the last terms, but humans might be unable to weight terms accordingly when interpreting a topic.

4.3. Quality of topical document representation – quantitative evaluation

For the evaluation of topical document representation, binary classification tasks are trained for each of the nine labels. For that matter, the document-topic probabilities θ_d of each model are used as independent variables to predict the manually given labels (dependent variable) for each response. This approach facilitates examining whether the topic models contain enough information to assign each response to the correct manual labels. Many algorithms such as logistic regression are available for training a binary classifier. For this study, Support Vector Machines (SVM) are chosen as they have shown to be very effective for text classification tasks (Manning et al., 2009). To compare the topic models, the F1 score is used, which a common metric to evaluate information retrieval (Van Rijsbergen, 1979). It measures how accurate the classifier predicts the positive cases, i.e., the cases where the manual label was assigned to a response (Manning et al., 2009). First, the F1 score is computed per classification task, i.e., per label, and then averaged over all labels to get one overall score for each topic model. Figure 9 gives an overview of the average F1 scores produced by the four methods. The scores for all models are found in Appendix B.

The lines in the figure depict the best F1 score reached by each method. It shows that LDA achieves the lowest scores for 80% of the data points. Moreover, at each data point there is at least one model that performs better than LDA. For $K \geq 15$, WNTM achieves the highest scores and its advantage over the other methods mainly increases with K. Aside from the method comparison, the graph shows that a higher number of topics leads overall to an increasing F1 score for all methods with few exceptions.

As the lines in Figure 9 only present the highest F1 scores achieved by each method, it can be questioned whether the superiority of WNTM depends on a certain hyperparameter setting. Hence, the shaded areas in the figure show the ranges of F1 scores for each method where the lower boundary indicates the lowest score achieved by a method and the upper boundary the highest one. The ranges achieved by BTM and WNTM are comparatively stable across all values of K, while LDA and LFLDA depend more strongly on the parameter setting. Hence, it cannot be deduced that the superiority of WNTM depends on a certain parameter setting. Moreover, there is no parameter setting for any method that always achieves the best performance.

So far, the F1 scores are averaged over all labels. However, as mentioned in chapter 2.1, some topic models like LDA struggle with topic imbalance, which often leads to the incapability to identify rare

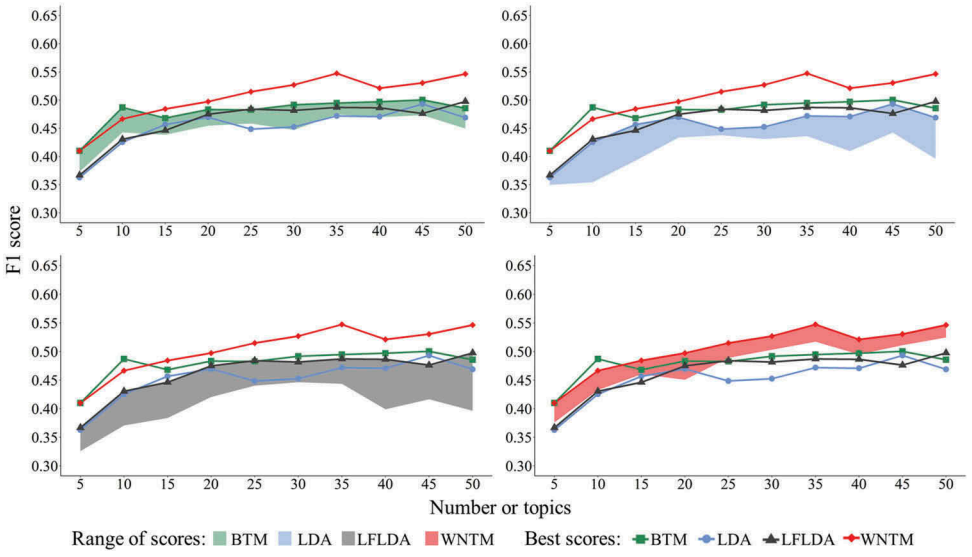

Figure 9. Best F1 score per method averaged over all labels (lines) and range of average F1 scores per method produced by different hyperparameter combinations (shaded area).

topics. As the label distribution in this study shows a notable imbalance (see Figure 3), we also investigated the methods' classification performance per label (see Appendix C). It can be observed that the scores differ remarkably between the labels and there seems to be a positive relation between the popularity of a label and the classification performance when predicting the same. For instance, a WNTM model achieves an F1 score of 0.7946 (best score across all values of K) for the label "Usability" which occurs in 30.19% of the responses. Meanwhile, the best F1 score achieved by WNTM for the label "Login", which is mentioned by 5.08% of the respondents, is only 0.5409. The same trend can be observed across all methods.

Altogether, WNTM achieves the best classification performance in terms of F1 score in most cases. When comparing the methods based on metrics that are averaged over all labels, one has to take into consideration that the classification performance differs notably between the nine labels. Overall, the labels that are frequently mentioned are predicted more accurately than the ones that are rarely mentioned.

4.4. Quality of topical document representation – qualitative evaluation

This section reports to which extent the topic distribution is consistent with the label distribution, regardless whether each document is assigned to the right topic or not.

The two methods considered for that are BTM with $\alpha^B = 0.05$ and $\beta = 0.1$ for $K = 10$ (F1 score: 0.4871) and WNTM with $\alpha = 0.1$ and $\beta = 0.1$ for $K = 20$ (F1 score: 0.4975). These are the models with the highest F1 scores for the respective values of K (see Appendix B).

Starting with BTM and $K = 10$, Table 6 presents the model's ten topics including top words. Each topic is assigned to one of the nine labels in coordination with two experts from the partner company. For most topics, the allocation is made only based on the top words while for a few topics that were less clear some top documents are considered to get more insights about the topics. Topic 10 cannot clearly be assigned to any label, even after reading some top documents. Further, no topics are available for the labels "Features", "Business" and "Data". In addition, there are some topics that seem to include two labels. For example, topic 1 entails words that indicate both labels "Usability" and "Documentation". However, based on the finding in section 0 that the topic-word probability drops significantly the later a word appears in the topic, more weight is put on the first words here. Therefore, topic 1 is assigned to "Usability".

Based on that allocation, Figure 10 shows the shares of documents that are assigned to each label via the document-topic distribution of the model mentioned above. The exact values are also depicted in Appendix D. Four different thresholds $t = \{0.18, 0.21, 0.24, 0.27\}$ to calculate the label distributions are reported here. Aside from the three labels mentioned above that are not present at all, the distributions derived via the thresholds differ in several points from the original label distribution. None of the thresholds leads to the same label distribution as the original one. Even when looking at single topics, there are only few relatively close matches. Regardless of the exact values, none of the thresholds leads to the correct ranking of labels that could reveal the relative importance of the topic compared to each other.

In the following, the same results are presented for the second exemplary model, namely WNTM and $K = 20$. Table 7 shows that this time each label is assigned to at least one topic.

The label distribution based on these topic allocations is depicted in Figure 11 (see Appendix D for the exact values). None of the thresholds $t = \{0.12, 0.15, 0.18, 0.21\}$ produces the same label

Table 6. Exemplary topics produced by BTM and corresponding labels.

Label	Topic	Top words
Usability	1	easy, use, api, well, documentation, good, platform, document, sdk, simple
Documentation	2	support, developer, time, good, well, problem, help, platform, community, sdk
	3	api, documentation, good, great, tool, lot, easy, graph, well, use
Satisfaction	4	app, platform, great, easy, good, ad, game, user, well, audience
	5	platform, developer, develop, like, recommend, use, work, can, good, api
Reach	6	people, user, can, easy, use, get, spread, know, find, way
	7	user, platform, reach, use, good, audience, lot, people, base, large
Must-have	8	developer, platform, use, web, app, media, develop, recommend, people, integrate
Login	9	user, login, app, use, easy, make, can, create, account, test
Features	*No topic*	
Business	*No topic*	
Data	*No topic*	
No label	10	page, platform, time, account, day, like, one, campaign, work

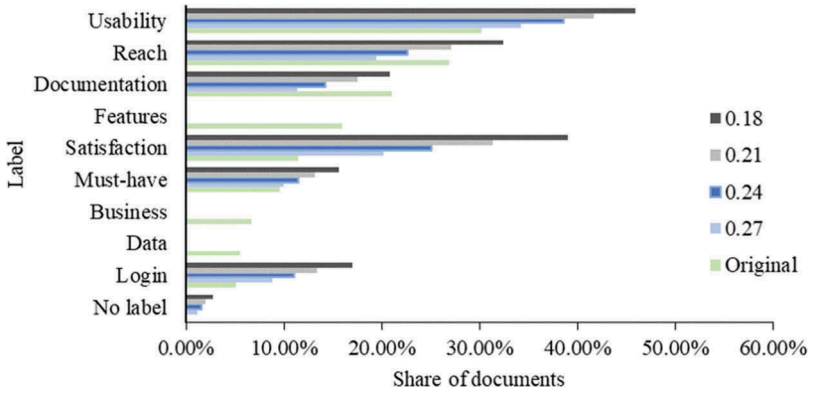

Figure 10. BTM topic distribution with different thresholds. Ordered by original share of labels.

Table 7. Exemplary topics produced by WNTM and corresponding labels.

Label	Topic	Top words
Usability	1	easy, api, use, well, document, documentation, simple, good, work, platform
	2	easy, get, understand, start, use, follow, simple, good, quick, documentation
	3	user, use, platform, easy, app, good, make, tool, develop, feature
Features	4	app, platform, mobile, game, tool, integration, sdk, feature, application, web
	5	api, graph, tool, test, developer, explorer, easy, great, testing, create
	6	get, recommend, can, platform, user, take, page, ad, lot, reach
Documentation	7	documentation, good, api, lot, platform, support, use, great, developer, example
	8	support, developer, time, response, issue, team, problem, help, good, bug
Satisfaction	9	great, platform, good, really, documentation, game, web, look, develop, give
	10	use, platform, recommend, develop, development, people, reason, many, ease, now
	11	time, use, lot, work, everything, people, can, much, thing, user
Data	12	user, datum, lot, use, access, information, get, login, can, integration
Reach	13	user, reach, audience, base, large, huge, people, platform, potential, wide
	14	people, user, get, can, way, easy, recommend, find, know, contact
	15	use, user, platform, can, make, need, people, account, almost, many
	16	good, ad, tool, way, app, great, platform, user, audience, target
Must-have	17	developer, platform, web, media, great, everyone, use, one, largest, popular
Login	18	login, user, app, use, easy, sign, create, share, web, account
Business	19	business, good, platform, can, great, develop, tool, lot, developer, many
No label	20	api, better, change, new, time, experience, improve, last, good, update

distributions as the original labelling. Yet, for many labels the approximation achieved by $t = 0.15$ is close to the original shares. Compared to Figure 10, the ranking of the labels is much better represented.

Altogether, this section provides insights on how well the topic distribution represents the original label distribution. First, the mapping of topics and labels demonstrates that in most cases the top words are sufficient to assign a topic to a label. In the remaining cases, the top documents provide further insights that facilitate the allocation. After that, the topic distributions are calculated for two models via different thresholds. The BTM solution with $K = 10$ does not cover all labels. In addition, the distribution of the covered labels differs significantly from the original one. However, the WNTM solution with $K = 20$ covers all labels and the label ranking is very similar to the original one, even if the exact distribution cannot be reproduced by any threshold.

5. Discussion

The use of text analytics is not yet considered an alternative to human coding, which has several reasonable grounds. First, an initial investment is required for tasks like finding the right algorithm and preparing the data before getting any insights. Further, topic modelling becomes significantly more accurate with an increasing number of responses (Tang et al., 2014) but a lot of market research studies suffer from a small amount of respondents (G. Lockot, personal communication, September, 2017). Content wise, topic modelling is inferior to the analysis through humans in several ways. Usually, topic models are not capable of discovering topics that are very detailed (Aggarwal & Zhai, 2012) or that show up rarely in the responses (Roberts et al., 2014). Moreover, it is easier to discover explicitly mentioned opinions with key words than implicitly described ones. While humans are mostly able to classify implicit mentions correctly based on common knowledge,

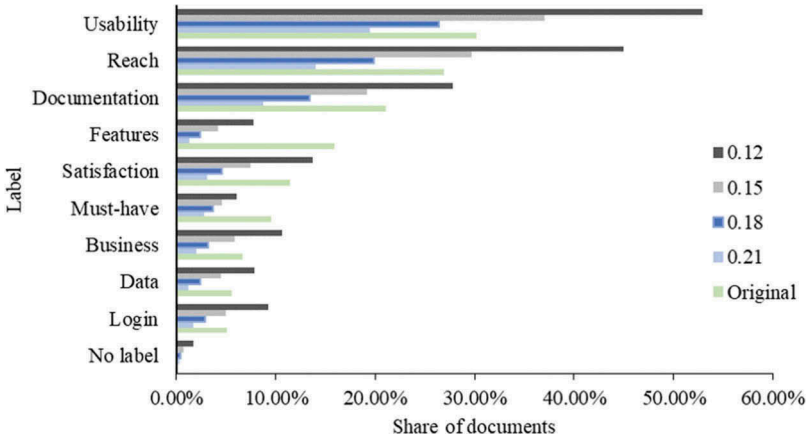

Figure 11. WNTM topic distribution with different thresholds. Ordered by original share of labels.

algorithms are usually not (Liu, 2012). Additionally, methods that are based on word co-occurrences and the "bag-of-words" assumption imply the limitation that semantics are ignored (Le & Mikolov, 2014). This makes the identification of implicit topics even more complicated as well as for example the identification of negation. Further, complex language like metaphors and humour are hard to analyse automatically without human intervention (Graesser & McNamara, 2012).

Still, topic modelling provides a lot of opportunities for the analysis of OE responses. The most obvious one is saving time and money (Roberts et al., 2014). Especially on a large scale, the up-front costs can pay off quickly. Moreover, topic modelling can also add value with regards to content. For instance, it facilitates the analysis of corpora where researchers cannot build upon any prior knowledge. Further, it can help to reduce several human biases. First, algorithms, in contrast to humans, identify topics objectively and do not assume them (Roberts et al., 2014). Second, algorithms provide consistency, which is a major drawback of human coding. It is well known that different human raters do not provide consistent results (between-rater variance) (Tinsley & Weiss, 1975). And even if all responses were analysed by the same researcher, there would still be inconsistencies as humans for instance get tired or bored (Graesser & McNamara, 2012).

The applicability of topic modelling was investigated from different perspectives in this work to gain an overall impression. The first part of the results was focused on whether the topics were clear enough to be used for exploratory purposes. A quantitative coherence score was used to compare the methods where BTM mostly achieved the best performance. To the best of the authors' knowledge, there is no absolute threshold though that differentiates a coherent from a non-coherent topic and therefore the metric is rather used for relative comparisons. However, it was shown that the ranking based on this metric was not always consistent with expert judgements. The results indicate that what makes up a high coherence score and what is perceived as clear and useful by researchers can be different. It can also be questioned whether the chosen coherence score is suitable for the present dataset. The fact that it only uses the target corpus is certainly advantageous in some respects. But the downside is that the calculation suffers from the lacking co-occurrence patterns of the corpus. Further, it has been shown that the interpretation of topics should focus on the first words in the top word lists. Overall, the experts from the partner company assessed the exemplary topics as clear and helpful.

The second part of the results focused on whether the responses were accurately represented by the topics, which was again investigated from two perspectives: First, document classification and respective metrics were used to explore whether the topics provided enough information to predict the right labels for each response. Second, it was examined to what extent the distribution of the original labels could be reproduced by using the topic distributions. WNTM achieved the highest classification performance in 80% of the cases. Yet, the performance achieved by all methods was only moderate (highest F1 score over all models: 0.5474, see Appendix B). It must be noted that the results indicate a relation between classification performance and the frequency with which a label is mentioned. The prediction is substantially more accurate for frequent than for rare topics. Even WNTM, for which the authors claim that it is capable of handling topic imbalance (Zuo et al., 2016), showed this relation.

The moderate performance on classification tasks does not imply that topic modelling is useless for the analysis of OE responses. Discussions with experts have confirmed that it was much more important to get a suitable topic distribution over the entire response set than a correct one-to-one mapping of responses and topics. Two exemplary models that showed comparatively good results on classification

were explored in that regard. For example, a WNTM model with 20 topics produced very promising results: All original labels and even important sub-labels could be identified. Although the original ranking of the labels could not be entirely reproduced, the big picture was correct aside from some exceptions.

Overall, it has been observed that the number of topics negatively affects the average topic quality and positively affects the average classification performance. However, it is assumed that a larger number of topics does not generally lead to less coherent topics. Rather it is plausible that only a limited number of topics is available to be identified and therefore the higher the number of topics, the higher the number of nonsense topics. Further, Yan et al. (2013) mention that a small number of topics usually leads to very general topics that are hard to distinguish while a larger number of topics produces more specific ones. To make sure that all relevant topics are identified, it is thus recommended for OE responses to choose a rather high number of topics and sort out the meaningless ones. In doing so, the researcher has the chance to recognize the small and specific topics and can still decide to combine them to a larger one.

In summary, the current work has shown that topic modelling bears high potential for the analysis of OE responses but does not provide a stand-alone solution. The experts from the partner company state that an automatic approach for exploration and a good approximation for the label ranking would already be a major gain for many studies. Certainly, this work only focused on one dataset and the opinion of experts from one company. An investigation on a larger scale would be interesting for future research.

Aside from the general usefulness of topic modelling for OE responses, this study's second focus is on the comparison of the four implemented methods. Through the implementation of short text topic models, it was possible to achieve better results than produced by the benchmark method LDA. BTM mainly achieved the best performance for topic coherence and WNTM for document classification. LFLDA produced very similar results to LDA and has the disadvantage that it depends on the availability and quality of external data. Finding a suitable external corpus is an additional effort required by LFLDA. While in this study the vocabulary is almost entirely represented by the chosen vector set, this could be an additional challenge for studies with a very domain-specific vocabulary. The studies that contributed the short text topic models considered here compare their respective innovation to LDA (Nguyen et al., 2015; Yan et al., 2013; Zuo et al., 2016). On the other hand, systematic comparisons of several short text topic models to one another are scarce. Therefore, our analysis of alternative short text topic models in the specific context of OE response processing expands the body of knowledge with original empirical evidence, which may be regarded as a more general contribution to the academic literature.

6. Conclusions

OE questions enjoy great popularity in market research studies but are associated with a very laborious and error-prone analysis called human coding. In this paper, we investigated the potential of four different topic models to be used as an alternative for human coding. Although it depends on the practical requirements whether topic modelling can replace the traditional approach, the study shows that topic models are very helpful for data exploration as well as topic ranking. Especially the dedicated short text topic models BTM (Yan et al., 2013) and WNTM (Zuo et al., 2016) achieve promising results. This provides a starting point for further research.

Notes

1. In the following, the *corpus size* refers to the number of documents.
2. To avoid misunderstanding, the corpus-topic distribution in BTM is labeled as α^B in the following.
3. In the following, method refers to the four topic modelling approaches implemented in this work (LDA, LFLDA, BTM, WNTM) while model refers to each fitted model instance of the methods with e.g., different hyperparameter settings.
4. The vector set is downloaded from https://nlp.stanford.edu/projects/glove/.
5. Note: The numbers in this figure do not add up to 100% as a document can be assigned to multiple labels.
6. Source code of LFLDA: https://github.com/datquocnguyen/LFTM. Source code of BTM: https://github.com/xiaohuiyan/BTM. Source code of WNTM: http://ipv6.nlsde.buaa.edu.cn/zuoyuan.
7. As a reminder: The experts do not know which topic belongs to which method.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The authors acknowledge financial support by the Open Access Publication Fund of the Humboldt-Universität zu Berlin.

References

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data* (1st ed.). New York: Springer.

Bicalho, P., Pita, M., Pedrosa, G., Lacerda, A., & Pappa, G. L. (2017). A general framework to expand short text for topic modeling. *Information Sciences*, *393*, 66–81.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brace, I. (2018). *Questionnaire design: How to plan, structure and write survey material for effective market research*. London: Kogan Page Publishers.

Brody, S., & Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 801–812). Stroudsburg, PA.: ACL.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, *22*, 288–296.

Converse, J. M., Jean McDonnell, C., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire* (Vol. 63). Thousand Oaks, CA: Sage.

Denny, M. J. (2017). SpeedReader: High performance text processing libraries. Retrieved from https://github.com/matthewjdenny/SpeedReader

Gendall, P., Menelaou, H., & Brennan, M. (1996). Open-ended questions: Some implications for mail survey research. *Marketing Bulletin*, *7*, 1–8.

Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper & P. Camic (Eds.), *APA handbook of research methods in psychology: Vol. 1. foundations, planning, measures, and psychometrics* (pp. 307–325). Washington, DC: American Psychological Association.

Griffiths, T. L., & Steyvers, M. (2001). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (p. 24). Fairfax, Virginia: George Mason University.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modelling in Twitter. In P. Melville (Ed.), *Proceedings of the first workshop on social media analytics* (pp. 80–88). New York, NY: ACM.

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*, 1–30.

Jin, O., Liu, N. N., Zhao, K., Yu, Y., & Yang, Q. (2011). Transferring topical knowledge from auxiliary long texts for short text clustering. In B. Berendt (Ed.), *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 775–784). New York, NY: ACM.

Kuhn, M. (2008). Caret package. *Journal of Statistical Software*, *28*, 1–26.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lau, H. J., Newman, D., & Baldwin, T. (2014). Machine Reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th conference of the European chapter of the association for computational linguistics* (pp. 530–539). Stroudsburg, PA: ACL.

Lazarsfeld, P. F. (1935). The art of asking WHY in marketing research: Three principles underlying the formulation of questionnaires. *National Marketing Review*, *1*, 26–38.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st international conference on machine learning* (pp. 1188–1196). Beijing, China.

Leleu, T. D., Jacobson, I. G., LeardMann, C. A., Smith, B., Foltz, P. W., Amoroso, P. J., & Smith, T. C. (2011). Application of latent semantic analysis for open-ended responses in a large, epidemiologic study. *BMC Medical Research Methodology*, *1–11*, 136.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*, 1–167.

Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Information Retrieval*, *14*, 178–203.

Manevitz, L. M., & Yousef, M. (2001). One-class SVMs for document classification. *Journal of Machine Learning Research*, *2*, 139–154.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to information retrieval*. Cambridge: Cambridge university press.

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In G. J. F. Jones (Ed.), *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (pp. 889–892). New York, NY: ACM.

Mihalcea, R., Courtney, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Aaai*, *6*, 775–780.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR workshop*. Scottsdale, Arizona.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 262–272). Stroudsburg, PA.: ACL.

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Searching microblogs: Coping with sparsity and document quality. In B. Berendt (Ed.), *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 183–188). New York, NY: ACM.

Newman, D., Karimi, S., & Cavedon, L. (2009). External evaluation of topic models. In J. Kay, P. Thomas, & A. Trotman (Eds.): *Vol. TR 645. Technical report/school of information technologies, University of Sydney, proceedings of the 14th Australasian document computing symposium* (pp. 11–18). Sydney: School of Information Technology, University of Sydney.

Newman, D., Lau, H. J., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In J. Baldridge, P. Clark, & G. Tur (Eds.), *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 100–108). Stroudsburg, PA: ACL.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, *3*, 299–313.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). Stroudsburg, PA: ACL.

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, & X. Zhang (Eds.), In *Proceeding of the 17th international conference on World Wide Web* (pp. 91–100). New York, NY: ACM.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*, 1064–1082.

Schouten, K., & Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *28*, 813–830.

Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *American Sociological Review*, *31*, 218–222.

Singh, V. K., Waila, P., Piryani, R., & Uddin, A. (2013). Computational exploration of theme-based blog data using topic modeling, NERC and sentiment classifier combine. *AASRI Procedia*, *4*, 212–222.

Sridhar, V. K. R. (2015). Unsupervised topic modeling for short texts using distributed representations of words. In *Proceedings of NAACL-HLT 2015* (pp. 192–200). Stroudsburg, PA: ACL.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. *Proceedings of the 31st International Conference on Machine Learning, PMLR 32*(1), 190–198.

Ten Kleij, F., & Musters, P. A. D. (2003). Text analysis of open-ended survey responses: A complementary method to preference mapping. *Food Quality and Preference*, *14*, 43–52.

Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*, 358–376.

Titov, I., & McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, & X. Zhang (Eds.), *Proceeding of the 17th international conference on World Wide Web* (pp. 111–120). New York, NY: ACM.

Tsai, F. S. (2011). A tag-topic model for blog mining. *Expert Systems with Applications*, *38*, 5330–5335.

Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In B. D. Davison (Ed.), *Proceedings of the third ACM international conference on Web search and data mining* (pp. 261–270). New York, NY: ACM.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. D. Schwabe, V. Almeida, & H. Glaser (Eds.), In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445–1456). New York, NY: ACM.

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In S. Macskassy (Ed.), *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 233–242). New York, NY: ACM.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee, & V. Murdock (Eds.): *Vol. 6611. Lecture notes in computer science, advances in information retrieval: 33rd European conference on IR resarch* (pp. 338–349). Berlin: Springer.

Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, *48*, 379–398.

Appendices

Appendix A. Label descriptions

The table provides a short description for each label.

Table A1. Short description of labels.

Label	Short description
Usability	The platform is easy to use
Reach	The platform reaches many users
Documentation	The documentation is easy to understand and the customer support is good
Features	The platform provides good features
Satisfaction	The platform is generally satisfactory without further specifying it
Must-have	The platform is widely accepted and thus inevitable
Business	Using the platform provides business opportunities
Data	The platform provides interesting insights
Login	The login process works well

Appendix B. Model performance metrics

The table provides the coherence score (averaged over all topic) and the F1 score (averaged over all labels) for all models fitted in this study.

Table B1. Coherence and F1 scores for BTM.

K	α/α^B	β	λ	Method	Coherence	F1
5	0.05	0.01	-	BTM	−113.4459	0.3732
10	0.05	0.01	-	BTM	−117.8001	0.4428
15	0.05	0.01	-	BTM	−121.3998	0.4515
20	0.05	0.01	-	BTM	−124.3161	0.4704
25	0.05	0.01	-	BTM	−124.7224	0.4586
30	0.05	0.01	-	BTM	−123.3720	0.4554
35	0.05	0.01	-	BTM	−124.0879	0.4816
40	0.05	0.01	-	BTM	−124.9988	0.4699
45	0.05	0.01	-	BTM	−127.7357	0.4733
50	0.05	0.01	-	BTM	−129.0318	0.4497
5	0.05	0.1	-	BTM	−115.6502	0.3787
10	0.05	0.1	-	BTM	−118.6618	0.4871
15	0.05	0.1	-	BTM	−116.1867	0.4681
20	0.05	0.1	-	BTM	−122.8126	0.4833
25	0.05	0.1	-	BTM	−123.9003	0.4741
30	0.05	0.1	-	BTM	−121.7328	0.4809
35	0.05	0.1	-	BTM	−126.9755	0.4949
40	0.05	0.1	-	BTM	−126.5275	0.4970
45	0.05	0.1	-	BTM	−123.9117	0.4859
50	0.05	0.1	-	BTM	−128.2998	0.4820
5	0.1	0.01	-	BTM	−112.3222	0.4101
10	0.1	0.01	-	BTM	−120.1509	0.4572
15	0.1	0.01	-	BTM	−118.3970	0.4494
20	0.1	0.01	-	BTM	−122.0090	0.4545
25	0.1	0.01	-	BTM	−119.0815	0.4768
30	0.1	0.01	-	BTM	−122.5052	0.4471
35	0.1	0.01	-	BTM	−126.2396	0.4746
40	0.1	0.01	-	BTM	−127.6996	0.4721
45	0.1	0.01	-	BTM	−128.1646	0.4791
50	0.1	0.01	-	BTM	−128.3404	0.4661
5	0.1	0.1	-	BTM	−118.8531	0.3737
10	0.1	0.1	-	BTM	−117.2850	0.4569
15	0.1	0.1	-	BTM	−121.8395	0.4388
20	0.1	0.1	-	BTM	−119.6477	0.4806
25	0.1	0.1	-	BTM	−121.6038	0.4827
30	0.1	0.1	-	BTM	−119.4750	0.4918
35	0.1	0.1	-	BTM	−127.0426	0.4900
40	0.1	0.1	-	BTM	−122.3340	0.4924
45	0.1	0.1	-	BTM	−126.5267	0.5004
50	0.1	0.1	-	BTM	−125.9982	0.4857

(Table of model performance metrics continued)

Table B2. Coherence and F1 scores for LDA.

	α/α^B	β	λ	Method	Coherence	F1
5	0.05	0.01	-	LDA	−113.5630	0.3497
10	0.05	0.01	-	LDA	−121.6064	0.3545
15	0.05	0.01	-	LDA	−125.9719	0.3927
20	0.05	0.01	-	LDA	−126.7626	0.4486
25	0.05	0.01	-	LDA	−129.9876	0.4449
30	0.05	0.01	-	LDA	−135.2573	0.4490
35	0.05	0.01	-	LDA	−134.3781	0.4719
40	0.05	0.01	-	LDA	−139.0463	0.4099
45	0.05	0.01	-	LDA	−141.1631	0.4788
50	0.05	0.01	-	LDA	−143.5997	0.3961
5	0.05	0.1	-	LDA	−112.6814	0.3564
10	0.05	0.1	-	LDA	−126.4861	0.4255
15	0.05	0.1	-	LDA	−122.0373	0.4271
20	0.05	0.1	-	LDA	−128.2056	0.4336
25	0.05	0.1	-	LDA	−128.8650	0.4378
30	0.05	0.1	-	LDA	−129.2903	0.4470
35	0.05	0.1	-	LDA	−132.0548	0.4391
40	0.05	0.1	-	LDA	−136.1988	0.4569
45	0.05	0.1	-	LDA	−134.9802	0.4425
50	0.05	0.1	-	LDA	−137.6962	0.4483
5	0.1	0.01	-	LDA	−118.9428	0.3630
10	0.1	0.01	-	LDA	−125.3804	0.4132
15	0.1	0.01	-	LDA	−129.1284	0.4316
20	0.1	0.01	-	LDA	−130.1269	0.4696
25	0.1	0.01	-	LDA	−137.1236	0.4486
30	0.1	0.01	-	LDA	−137.3304	0.4310
35	0.1	0.01	-	LDA	−139.5810	0.4363
40	0.1	0.01	-	LDA	−144.2553	0.4707
45	0.1	0.01	-	LDA	−147.3176	0.4932
50	0.1	0.01	-	LDA	−146.9439	0.4689
5	0.1	0.1	-	LDA	−112.3460	0.3585
10	0.1	0.1	-	LDA	−119.2432	0.4090
15	0.1	0.1	-	LDA	−128.2986	0.4568
20	0.1	0.1	-	LDA	−131.1908	0.4405
25	0.1	0.1	-	LDA	−131.8484	0.4472
30	0.1	0.1	-	LDA	−133.2490	0.4525
35	0.1	0.1	-	LDA	−139.3548	0.4695
40	0.1	0.1	-	LDA	−140.5986	0.4571
45	0.1	0.1	-	LDA	−142.6182	0.4544
50	0.1	0.1	-	LDA	−144.0240	0.4486

(Table of model performance metrics continued)

Table B3. Coherence and F1 scores for LFLDA.

K	α/α^B	β	λ	Method	Coherence	F1
5	0.05	0.01	0.6	LFLDA	−113.7867	0.3659
10	0.05	0.01	0.6	LFLDA	−119.1345	0.3709
15	0.05	0.01	0.6	LFLDA	−123.5224	0.4267
20	0.05	0.01	0.6	LFLDA	−126.0736	0.4313
25	0.05	0.01	0.6	LFLDA	−127.4000	0.4404
30	0.05	0.01	0.6	LFLDA	−132.7093	0.4559
35	0.05	0.01	0.6	LFLDA	−134.1155	0.4541
40	0.05	0.01	0.6	LFLDA	−138.9572	0.3991
45	0.05	0.01	0.6	LFLDA	−139.8065	0.4165
50	0.05	0.01	0.6	LFLDA	−140.1242	0.4797
5	0.05	0.01	1	LFLDA	−115.0354	0.3425
10	0.05	0.01	1	LFLDA	−117.2286	0.4306
15	0.05	0.01	1	LFLDA	−124.1351	0.4200
20	0.05	0.01	1	LFLDA	−128.7456	0.4400
25	0.05	0.01	1	LFLDA	−130.5739	0.4497
30	0.05	0.01	1	LFLDA	−133.0834	0.4465
35	0.05	0.01	1	LFLDA	−136.0601	0.4437
40	0.05	0.01	1	LFLDA	−137.7121	0.4554
45	0.05	0.01	1	LFLDA	−140.4156	0.4542
50	0.05	0.01	1	LFLDA	−141.1532	0.4760
5	0.05	0.1	0.6	LFLDA	−121.5354	0.3669
10	0.05	0.1	0.6	LFLDA	−119.3704	0.3781
15	0.05	0.1	0.6	LFLDA	−122.6553	0.3839
20	0.05	0.1	0.6	LFLDA	−125.2916	0.4204
25	0.05	0.1	0.6	LFLDA	−127.6703	0.4420
30	0.05	0.1	0.6	LFLDA	−131.8856	0.4466
35	0.05	0.1	0.6	LFLDA	−134.0838	0.4495
40	0.05	0.1	0.6	LFLDA	−138.2579	0.4494
45	0.05	0.1	0.6	LFLDA	−138.7113	0.4450
50	0.05	0.1	0.6	LFLDA	−139.6516	0.4267
5	0.05	0.1	1	LFLDA	−115.1217	0.3355
10	0.05	0.1	1	LFLDA	−121.2215	0.4240
15	0.05	0.1	1	LFLDA	−122.4510	0.4205
20	0.05	0.1	1	LFLDA	−128.5243	0.4315
25	0.05	0.1	1	LFLDA	−128.1921	0.4601
30	0.05	0.1	1	LFLDA	−131.0169	0.4589
35	0.05	0.1	1	LFLDA	−131.7854	0.4537
40	0.05	0.1	1	LFLDA	−136.8598	0.4675
45	0.05	0.1	1	LFLDA	−137.2851	0.4475
50	0.05	0.1	1	LFLDA	−138.3750	0.3964

(Table of model performance metrics continued)

Table B4. Coherence and F1 scores for LFLDA (continued).

K	α/α^B	β	λ	Method	Coherence	F1
5	0.1	0.01	0.6	LFLDA	−117.6724	0.3512
10	0.1	0.01	0.6	LFLDA	−121.8220	0.4181
15	0.1	0.01	0.6	LFLDA	−126.5715	0.4385
20	0.1	0.01	0.6	LFLDA	−131.8531	0.4267
25	0.1	0.01	0.6	LFLDA	−134.9967	0.4841
30	0.1	0.01	0.6	LFLDA	−139.3103	0.4493
35	0.1	0.01	0.6	LFLDA	−139.5290	0.4768
40	0.1	0.01	0.6	LFLDA	−140.7336	0.4816
45	0.1	0.01	0.6	LFLDA	−146.9044	0.4602
50	0.1	0.01	0.6	LFLDA	−147.8339	0.4640
5	0.1	0.01	1	LFLDA	−114.7362	0.3259
10	0.1	0.01	1	LFLDA	−123.9614	0.4208
15	0.1	0.01	1	LFLDA	−125.9456	0.4328
20	0.1	0.01	1	LFLDA	−131.1523	0.4574
25	0.1	0.01	1	LFLDA	−132.5545	0.4803
30	0.1	0.01	1	LFLDA	−138.8816	0.4817
35	0.1	0.01	1	LFLDA	−141.2121	0.4818
40	0.1	0.01	1	LFLDA	−144.6030	0.4865
45	0.1	0.01	1	LFLDA	−146.6254	0.4751
50	0.1	0.01	1	LFLDA	−147.0708	0.4976
5	0.1	0.1	0.6	LFLDA	−117.6574	0.3547
10	0.1	0.1	0.6	LFLDA	−125.1680	0.3745
15	0.1	0.1	0.6	LFLDA	−126.6801	0.4316
20	0.1	0.1	0.6	LFLDA	−130.5593	0.4480
25	0.1	0.1	0.6	LFLDA	−133.1574	0.4525
30	0.1	0.1	0.6	LFLDA	−135.3356	0.4520
35	0.1	0.1	0.6	LFLDA	−139.5369	0.4611
40	0.1	0.1	0.6	LFLDA	−141.1558	0.4583
45	0.1	0.1	0.6	LFLDA	−146.5139	0.4472
50	0.1	0.1	0.6	LFLDA	−145.9266	0.4482
5	0.1	0.1	1	LFLDA	−115.3304	0.3556
10	0.1	0.1	1	LFLDA	−121.9343	0.4192
15	0.1	0.1	1	LFLDA	−126.3659	0.4463
20	0.1	0.1	1	LFLDA	−133.1316	0.4751
25	0.1	0.1	1	LFLDA	−134.2372	0.4610
30	0.1	0.1	1	LFLDA	−134.0476	0.4775
35	0.1	0.1	1	LFLDA	−138.2616	0.4872
40	0.1	0.1	1	LFLDA	−141.7594	0.4450
45	0.1	0.1	1	LFLDA	−144.9326	0.4762
50	0.1	0.1	1	LFLDA	−144.7627	0.4833

(Table of model performance metrics continued)

Table B5. Coherence and F1 scores for WNTM.

K	α/α^{β}	β	λ	Method	Coherence	F1
5	0.05	0.01	-	WNTM	−119.4465	0.4039
10	0.05	0.01	-	WNTM	−123.8929	0.4666
15	0.05	0.01	-	WNTM	−129.8543	0.4665
20	0.05	0.01	-	WNTM	−138.7321	0.4508
25	0.05	0.01	-	WNTM	−134.9821	0.4890
30	0.05	0.01	-	WNTM	−135.3235	0.5033
35	0.05	0.01	-	WNTM	−140.3908	0.5191
40	0.05	0.01	-	WNTM	−139.2001	0.5075
45	0.05	0.01	-	WNTM	−141.1103	0.5262
50	0.05	0.01	-	WNTM	−143.1343	0.5300
5	0.05	0.1	-	WNTM	−119.2546	0.3926
10	0.05	0.1	-	WNTM	−122.5133	0.4327
15	0.05	0.1	-	WNTM	−119.1964	0.4844
20	0.05	0.1	-	WNTM	−130.8497	0.4901
25	0.05	0.1	-	WNTM	−130.3037	0.4929
30	0.05	0.1	-	WNTM	−132.9228	0.5191
35	0.05	0.1	-	WNTM	−136.1350	0.5173
40	0.05	0.1	-	WNTM	−137.3759	0.5212
45	0.05	0.1	-	WNTM	−136.1579	0.5305
50	0.05	0.1	-	WNTM	−140.2123	0.5277
5	0.1	0.01	-	WNTM	−121.9704	0.3761
10	0.1	0.01	-	WNTM	−128.2620	0.4560
15	0.1	0.01	-	WNTM	−130.4949	0.4705
20	0.1	0.01	-	WNTM	−130.6631	0.4945
25	0.1	0.01	-	WNTM	−132.6220	0.4972
30	0.1	0.01	-	WNTM	−137.0506	0.5269
35	0.1	0.01	-	WNTM	−139.4586	0.5247
40	0.1	0.01	-	WNTM	−141.1178	0.5044
45	0.1	0.01	-	WNTM	−142.5379	0.5189
50	0.1	0.01	-	WNTM	−143.5525	0.5246
5	0.1	0.1	-	WNTM	−118.1253	0.4100
10	0.1	0.1	-	WNTM	−120.9275	0.4582
15	0.1	0.1	-	WNTM	−123.3076	0.4603
20	0.1	0.1	-	WNTM	−127.4511	0.4975
25	0.1	0.1	-	WNTM	−127.7618	0.5148
30	0.1	0.1	-	WNTM	−133.3650	0.5238
35	0.1	0.1	-	WNTM	−131.3758	0.5474
40	0.1	0.1	-	WNTM	−137.1212	0.4950
45	0.1	0.1	-	WNTM	−139.4415	0.5110
50	0.1	0.1	-	WNTM	−140.0693	0.5465

Appendix C. F1 Score per label

The table shows the best F1 score per method (over all hyperparameter settings) for each label.

Table C1. Best F1 scores per method for each label.

	Method	Labels								
K		Usability	Reach		Features	Satis-faction	Must-have	Business	Data	Login
Share of responses		30.19%	26.89%	21.02%	15.92%	11.44%	9.56%	6.70%	5.54%	5.08%
5	LFLDA	0.5981	0.6031	0.6164	0.3556	0.2662	0.4014	0.3448	0.1913	0.3248
10	LFLDA	0.6557	0.6367	0.5878	0.4078	0.3175	0.4659	0.3745	0.2140	0.4142
15	LFLDA	0.6528	0.6370	0.6072	0.4602	0.3194	0.4719	0.3636	0.2410	0.4432
20	LFLDA	0.7036	0.6197	0.6346	0.4186	0.3403	0.5018	0.4188	0.3128	0.4557
25	LFLDA	0.7129	0.6287	0.6517	0.4720	0.3506	0.4375	0.4036	0.3750	0.4857
30	LFLDA	0.7204	0.6409	0.6498	0.4772	0.3533	0.4416	0.4145	0.3377	0.4233
35	LFLDA	0.7365	0.6527	0.6622	0.4810	0.3740	0.4377	0.4167	0.3350	0.4906
40	LFLDA	0.7193	0.6553	0.6623	0.4620	0.3722	0.4527	0.4278	0.3333	0.4780
45	LFLDA	0.7279	0.6512	0.6681	0.4279	0.3626	0.4291	0.4072	0.3265	0.5170
50	LFLDA	0.7358	0.6503	0.7029	0.4727	0.3537	0.4490	0.3881	0.4096	0.4691
5	WNTM	0.6290	0.6331	0.5936	0.4082	0.3182	0.4476	0.2701	0.2389	0.3482
10	WNTM	0.6873	0.6792	0.6905	0.4762	0.3832	0.4690	0.3917	0.2168	0.3744
15	WNTM	0.6869	0.6752	0.6855	0.4332	0.3911	0.4859	0.3745	0.3130	0.4211
20	WNTM	0.7492	0.6704	0.6923	0.4307	0.4419	0.4730	0.3843	0.3033	0.4706
25	WNTM	0.7384	0.6813	0.7048	0.5063	0.4524	0.4689	0.4066	0.3239	0.4598
30	WNTM	0.7783	0.6728	0.7273	0.4965	0.4534	0.4769	0.4192	0.3448	0.4778
35	WNTM	0.7584	0.6984	0.7350	0.4844	0.4922	0.4926	0.4653	0.3347	0.5409
40	WNTM	0.7793	0.6972	0.7137	0.4711	0.4363	0.4585	0.4158	0.4095	0.4494
45	WNTM	0.7726	0.6976	0.7636	0.5316	0.4498	0.4912	0.4231	0.3905	0.4686
50	WNTM	0.7946	0.6923	0.7533	0.5347	0.4698	0.4681	0.4093	0.3474	0.5161
5	LDA	0.5911	0.6104	0.5366	0.3306	0.2662	0.2903	0.2598	0.2431	0.3258
10	LDA	0.6748	0.6188	0.5804	0.3446	0.2805	0.4000	0.3825	0.2414	0.4088
15	LDA	0.6699	0.6565	0.5708	0.4152	0.3158	0.4206	0.3645	0.3304	0.4286
20	LDA	0.7043	0.6567	0.5885	0.3747	0.3558	0.4268	0.3614	0.3493	0.4578
25	LDA	0.7190	0.6446	0.6233	0.4023	0.3437	0.4138	0.3806	0.2832	0.4000
30	LDA	0.7351	0.6436	0.6441	0.3936	0.3456	0.3860	0.3664	0.3770	0.3949
35	LDA	0.7608	0.6692	0.6376	0.4071	0.3491	0.3604	0.4257	0.2807	0.4933
40	LDA	0.7521	0.6728	0.6590	0.4362	0.3284	0.3894	0.3846	0.3128	0.4551
45	LDA	0.7608	0.6705	0.6711	0.4255	0.3443	0.4271	0.4103	0.3830	0.4487
50	LDA	0.7568	0.6441	0.6374	0.4368	0.3351	0.3958	0.4206	0.3053	0.4162

(Table of F1 score per label continued)

Table C2. Best F1 scores per method for each label (continued).

	Method	Labels								
K		Usability	Reach		Features	Satis-faction	Must-have	Business	Data	Login
5	BTM	0.5837	0.6234	0.6008	0.3045	0.3034	0.5061	0.2638	0.2206	0.3842
10	BTM	0.6429	0.6667	0.6450	0.4303	0.3642	0.5136	0.3898	0.2366	0.4324
15	BTM	0.6655	0.6753	0.6462	0.4167	0.3793	0.4715	0.4018	0.2340	0.4379
20	BTM	0.7057	0.6716	0.6624	0.4478	0.3963	0.4882	0.4279	0.2385	0.4649
25	BTM	0.6974	0.6728	0.6796	0.4513	0.3893	0.4752	0.4105	0.2929	0.4432
30	BTM	0.7013	0.6716	0.6594	0.4471	0.3977	0.4806	0.4370	0.2689	0.4422
35	BTM	0.7013	0.6846	0.6580	0.4737	0.4498	0.4788	0.4167	0.2857	0.4734
40	BTM	0.6966	0.6704	0.6609	0.4366	0.4333	0.4828	0.3665	0.3723	0.4663
45	BTM	0.7097	0.6654	0.6814	0.4326	0.3920	0.4876	0.4151	0.3148	0.4678
50	BTM	0.7201	0.6679	0.6538	0.4193	0.4204	0.4621	0.3863	0.3370	0.4713

Appendix D. Topic distribution with different thresholds

The following table provides the topic distribution for the BTM model with $K = 10$, $\alpha^B = 0.05$ and $\beta = 0.1$ after matching the topics with the original labels.

Table D1. Topic distribution for selected BTM model.

		Threshold-based label share (with threshold t)			
Label	Original label share	$t = 0.27$	$t = 0.24$	$t = 0.21$	$t = 0.18$
Usability	30.19%	34.28%	38.61%	41.68%	45.88%
Reach	26.89%	19.47%	22.62%	27.07%	32.47%
Documentation	21.02%	11.39%	14.25%	17.50%	20.80%
Features	15.92%	0.00%	0.00%	0.00%	0.00%
Satisfaction	11.44%	20.18%	25.11%	31.32%	38.97%
Must-have	9.56%	10.00%	11.43%	13.15%	15.64%
Business	6.70%	0.00%	0.00%	0.00%	0.00%
Data	5.54%	0.00%	0.00%	0.00%	0.00%
Login	5.08%	8.79%	11.07%	13.41%	16.99%
No label	0.00%	1.17%	1.57%	2.06%	2.72%

The following table provides the topic distribution for the WNTM model with $K = 20$, $\alpha = 0.1$ and $\beta = 0.1$ after matching the topics with the original labels.

Table D2. Topic distribution for selected WNTM model.

		Threshold-based label share (with threshold t)			
Label	Original label share	$t = 0.21$	$t = 0.18$	$t = 0.15$	$t = 0.12$
Usability	30.19%	19.49%	26.41%	37.02%	52.91%
Reach	26.89%	14.01%	19.86%	29.71%	45.00%
Documentation	21.02%	8.71%	13.43%	19.17%	27.80%
Features	15.92%	1.27%	2.39%	4.15%	7.81%
Satisfaction	11.44%	3.06%	4.57%	7.42%	13.68%
Must-have	9.56%	2.75%	3.66%	4.55%	6.06%
Business	6.70%	2.05%	3.15%	5.89%	10.59%
Data	5.54%	1.21%	2.40%	4.47%	7.88%
Login	5.08%	1.68%	2.90%	4.97%	9.22%
No label	0.00%	0.29%	0.41%	0.70%	1.73%